# Optimizing Groundwater Conditioning Parameters for Groundwater Potential Identification Using Machine Learning Approaches in Klang and Langat River Basins

Syed Zabidi S.R.[1], Bohari S.N.[1*], Saian R.[2], Haron Narashid R.[1]

[1]Faculty of Built Environment, Surveying Science and Geomatics Studies, Universiti Teknologi MARA, Perlis Branch, Arau Campus, Malaysia

[2]Faculty of Computer and Mathematical Sciences, Surveying Science and Geomatics Studies, Universiti Teknologi MARA, Perlis Branch, Arau Campus, Malaysia

*ashikin10@uitm.edu.my

***Abstract:*** *Groundwater potential (GWP) studies relied heavily on the appropriate selection of parameters. Past studies have considered factors such as topography, hydrology, geology, land cover and climate changes; however, not all variables contribute equally to groundwater occurrence. Therefore, the proper selection of parameters is essential to ensure the accuracy and reliability of GWP prediction. This study aims to optimize 20 GWP conditioning parameters by utilizing several statistical approaches: correlation matrix, multicollinearity and chi-square tests. These parameters include elevation, slope, aspect, curvature, plan curvature, profile curvature, topographical wetness index (TWI), topographical roughness index (TRI), topographical position index (TPI) and stream position index (SPI), drainage density, lineament density, geology, lithology, distance to fault, distance to streams and distance to main tributaries, soil types, normalized difference vegetation index (NDVI), land use land cover (LULC) and rainfall. The correlation analysis and multicollinearity test results indicate all parameters fall within the required thresholds, showing minimal redundancy and no multicollinearity issues. In contrast, the results from chi-square test indicate that 9 parameters: lineament density, elevation, geology, soil, slope, distance to fault, LULC, NDVI and drainage density exhibit significant contribution (p-value<0.05) and therefore are retained for GWP prediction. These optimized parameters were then applied to predict GWP areas in the Klang and Langat River basins using the random forest (RF) machine learning technique. 564 tubewell points were allocated with 70% for training and 30% for testing. The findings indicated that the areas with the highest groundwater potential were located in the middle part of the basins, with a percentage of 14.02%. In contrast, the areas with the lowest groundwater potential were located in the northern and northeastern areas, with the percentages of 20.58%. The evaluation indicates the model exhibited strong performance, achieving an area under the curve (AUC) value of 0.927 for training and 0.860 for testing. The findings of this study will improve the precision and dependability of groundwater potential mapping for future sustainable groundwater management systems.*

*Keywords: Groundwater potential identification, optimization, machine learning, random forest*

## Introduction

Groundwater is a critical natural resource, serving as a primary supply of fresh water. Rapid population growth, urbanization, and land development have led to an increasing global demand for groundwater. In Malaysia, several states, such as Kelantan, Perlis, and

Negeri Sembilan have employed groundwater as a substitute resource to address the problem of water shortage (Azizan et al., 2018; M. M. A. Khan et al., 2021). Consequently, the over extraction of groundwater resources may result in substantial issues, including land subsidence, polluted water, and aquifer deterioration (El Shinawi et al., 2022; Panneerselvam et al., 2023; Sharifi et al., 2024). Therefore, identifying groundwater potential resources is essential as it is the first step towards efficient groundwater management and ensuring long-term sustainability.

Numerous past researchers have identified GWP by using various method ranging from conventional to advanced machine learning techniques, in order to achieve accurate and reliable outcomes (Das & Saha, 2022; Prasad et al., 2020). A careful selection of relevant conditioning parameters is essential to ensure the effectiveness of the method (Fatah et al., 2024). Typically, most of the past researchers have selected GWP conditioning parameters based on various factors such as topography, hydrology, geology, land cover, and climate changes that contribute to the movement and occurrence of the groundwater resources (Fatah et al., 2024; Jari et al., 2023; Liu et al., 2022; Prasad et al., 2020). Therefore, a systematic selection and evaluation of the appropriate parameters is significant to enhance the reliability and accuracy of the GWP outcome.

The identification of groundwater is complex, considering a large number of parameters to effectively implement the model. However, some of the parameters might not significantly influence the GWP results, leading to redundancy and multicollinearity issues (Liu et al., 2022; Ouali et al., 2023). Therefore, to resolve this issue, several statistical approaches such as correlation analysis, multicollinearity test and chi-square test have been employed to optimize and enhance the selection of parameters. Studies from Arabameri et al. (2019), Ghosh & Bera (2024), Moghaddam et al. (2020) and Sharma et al. (2024) demonstrates that these methods successfully identified significant relationships and emphasized the most influential GWP parameters. The variables with no predictive ability must be eliminated to ensure a reliable and accurate GWP results (Liu et al., 2022).

Thus, the optimization of the selected parameters is essential before utilizing the GWP identification (Kalantar et al., 2019; Ouali et al., 2023). This study aims to optimize and analyze the parameters derived from factors of topography, hydrogeology, land covers, and climate changes, using correlation matrix, multicollinearity test, and chi-square test. The variables that were not within required threshold were eliminated to retain the most relevant variables for GWP modelling. Then, the identification of the GWP area was conducted by

using one of the commonly used machine learning methods, which is random forest (RF). The evaluations were utilized by using several statistical metrics. In summary, this study highlights the significance of parameter optimization and the reliability of RF in producing accurate and reliable GWP results.

## Study Area

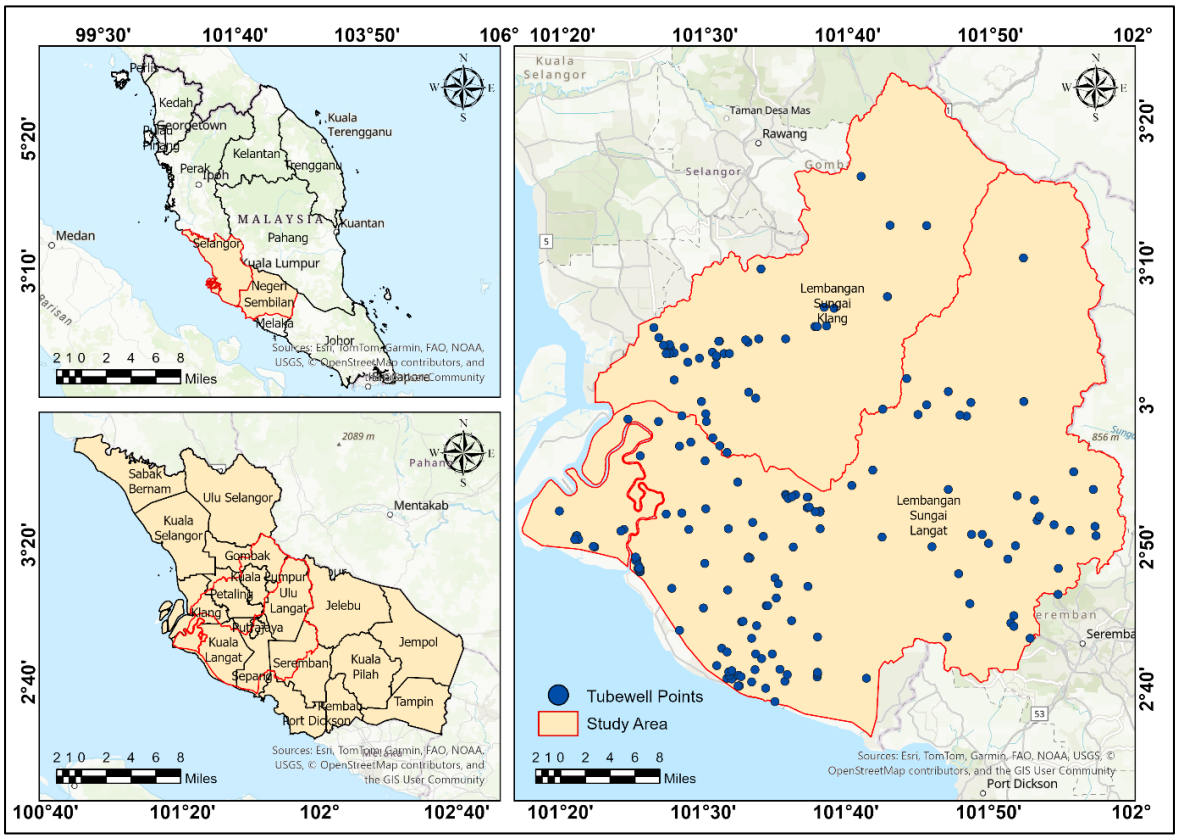


Figure 1: Study Area

Figure 1 illustrates the study area encompasses two different basins in Selangor and Kuala Lumpur: the Klang and Langat Rivers. Klang River basin is located on the western coast of Peninsular Malaysia, with coordinates of 3˚ 00' 0.00" N and 101° 22' 59.99" E. The river basin area is approximately 1288 km$^2$, which includes a majority of Kuala Lumpur state with several districts of Selangor, which are half of Gombak and Klang, the majority of Petaling, and a small portion of Sepang. Meanwhile, the Langat River basin is located in the southern part of Selangor, with the latitude of 2˚ 40' 0.00" N to 3˚ 20' 0.00" N and longitude of 101° 10' 00" E to 102° 00' 00" E. The river basin area is estimated to be 2432 km$^2$, with a total length of 183.65 km. This basin includes a majority of Putrajaya states and several Selangor districts, including Kuala Langat, Hulu Langat, Sepang, a small portion of Kuala Lumpur, and a part of Negeri Sembilan.

## Methodology

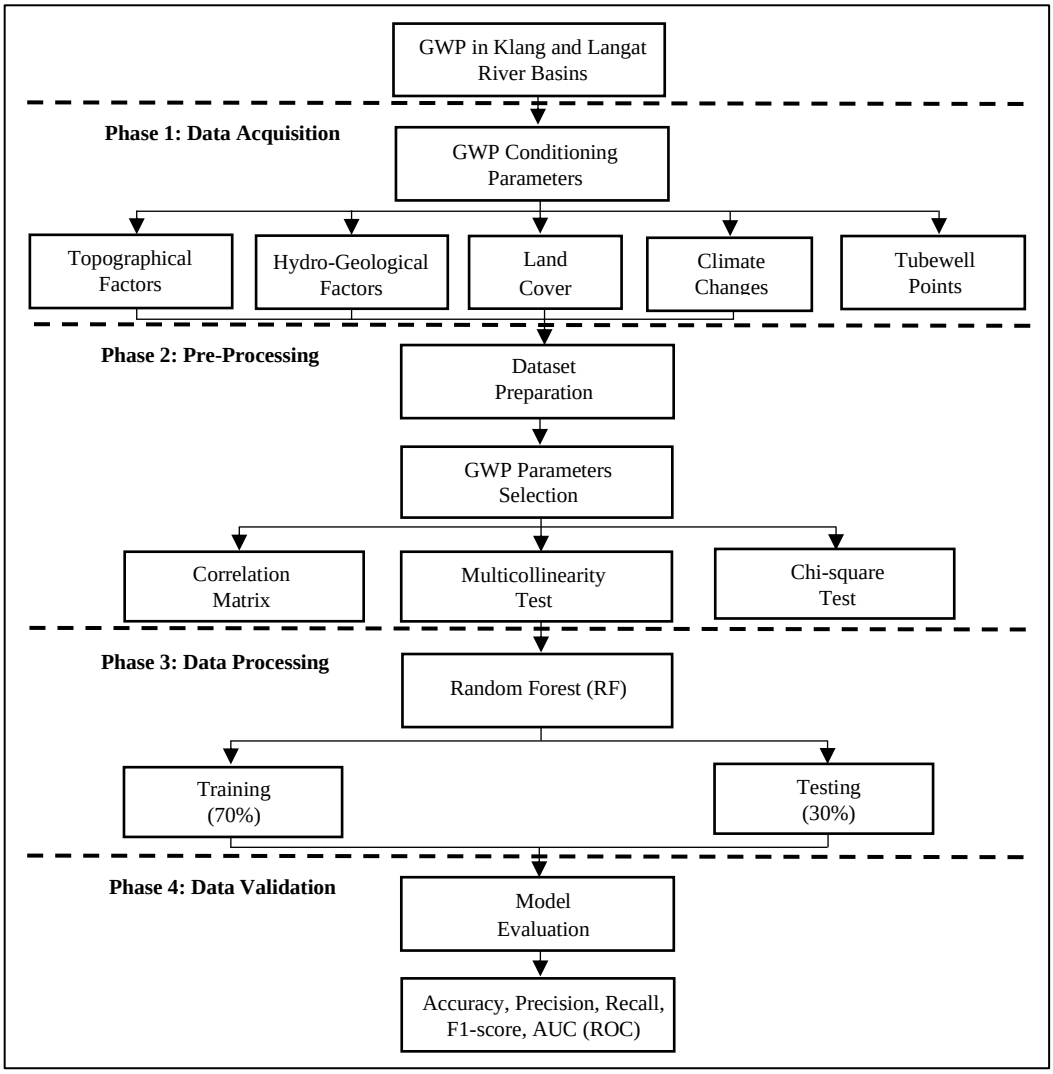


Figure 2: Methodology Flow Chart

Figure 2 illustrates the methodological workflow of this study, including four main phases: data acquisition, pre-processing, data processing, and data validation. The preparation began with the preparation of the thematic layers for each GWP conditioning parameters using ArcGIS software. These layers were then extracted at sample location points for dataset preparation. The parameter selection was carried out in the pre-processing stages using the correlation matrix, multicollinearity, and chi-square tests. These processes ensure no redundancy among the variables and can retain the most valuable factors. The RF was then applied to predict the GWP area with the dataset partitioned between 70% for training and 30% for testing. Lastly, the model's performance was assessed using accuracy, precision, recall, AUC, and ROC curve metrics for both training and testing datasets.

### a. Training and Testing Dataset

In this study, the training and testing dataset were consists of 564 tubewell points and 564 non-tubewell points that were randomly generated across study area. The tubewell points were assigned as 1 and non-tubewell points were assigned as 0. The dataset then divided into 70% (training) and 30% (testing) in order to predict the groundwater potential by using RF. Figure 1 depicts the tubewell distribution map across study area.

### b. Groundwater Conditioning Parameters

There were total of 20 groundwater conditioning parameters that gives influence towards groundwater potential studies. These parameters have been divided into four different factors which are topography (elevation, slope, aspect, plan curvature, profile curvature, TWI, TRI, TPI, SPI), hydrogeology (drainage density, lineament density, geology, lithology, distance to fault, distance to stream, distance to main tributaries), land cover (LULC, soil types, NDVI) and climate changes (rainfall).

#### i. Elevation

Elevation affects groundwater as water naturally flows on low topography compared to high topography (Anh et al., 2023; Thanh et al., 2022). Consequently, high elevation indicates low groundwater potential, while low elevation indicates high groundwater potential (Masroor et al., 2021; Moghaddam et al., 2020; Wei et al., 2022). Figure 3(a) illustrates the elevation map of research region, with altitudes varying from –118m to 1412m.

#### ii. Slope

Slope influences groundwater infiltration, surface runoff and groundwater movement (Al-Kindi & Janizadeh, 2022; Razavi-Termeh et al., 2024). Gentle slopes have an elevated infiltration rate, enhancing water capacity (Fatah et al., 2024; Islam et al., 2023; P. Kumar et al., 2024; Shandu & Atif, 2023). Meanwhile, steep slopes have lower infiltration, reducing water capacity (Al-Kindi & Janizadeh, 2022; Islam et al., 2023; Z. A. Khan & Jhamnani, 2023; Shandu & Atif, 2023). Figure 3(b) illustrates a slope map made using Slope tools that ranges from 0 to 75 degrees.

#### iii. Aspect

Aspect denotes the slope's direction and orientation, which is essential for groundwater occurrence (Fatah et al., 2024; Roy et al., 2024; Sharma et al., 2024). Figure 3(c) depicts the aspect map derived from the Aspect tool, categorized into ten directional groups.

### iv. Plan Curvature

Negative values on profile curvature signify a concave slope, resulting in the convergence of water flows, while positive values show a convex slope, reflecting the divergence of water flow over the surface (M. Kumar et al., 2023; Prasad et al., 2020). Figure 3(d) illustrates the plan curvature map categorized into convex, linear and concave types.

### v. Profile Curvature

In contrast to plan curvature, profile curvature indicates the acceleration and slowdown of water flow (R. Kumar et al., 2021). A negative profile curvature value signifies a deceleration in surface water flow, while a positive value denotes an acceleration in surface water flow (Ghosh & Bera, 2024; M. Kumar et al., 2023; Prasad et al., 2020). A value of 0 signifies a linear surface. The figure in 3(e) illustrates the profile curvature map of the Klang and Langat River basins.

### vi. TWI

TWI or topographical wetness index signifies the impact of topography on hydrological processes (Z. A. Khan & Jhamnani, 2023; Razavi-Termeh et al., 2024). It will determine the potential of the water flow or accumulation (Sørensen et al., 2006). The TWI can be calculated using the formula provided below (Moore et al., 1991).

$$TWI = In(\frac{fa}{tan\beta}) \quad (1)$$

The formula designates $fa$ as the flow accumulations and $\beta$ as the slope angle at the specified location (Prasad et al., 2020). Figure 3(f) illustrates the TWI map, with a range from 2 to 36.

### vii. TRI

Topographical Roughness Index (TRI) assesses the elevation difference between DEM cells to calculate terrain roughness and monitor the changes on the ground surface (Z. A. Khan & Jhamnani, 2023; Seifu et al., 2023). The TRI is determined using the formula provided below (Riley & Degloria, 1999).

$$TRI = (F_S mean - F_S min)/(F_S max - F_S min) \quad (2)$$

$F_S mean$ refers as mean focal statistic, while $F_S max$ signifies the highest focal statistics and $F_S min$ indicates the lowest focal statistic of a surface (Mukherjee & Singh, 2020). Figure 3(g) illustrates the TRI map with the values ranging from 0 to 1.

### viii. TPI

TPI or topographical position index, illustrated in Figure 3(h) measures the elevation of a target cell compared to its neighbouring cells (Dahal et al., 2023; Roy et al., 2024). The TPI can be derived using the formula below.

$$TPI = (\frac{E_P}{E_S}) \quad (3)$$

$E_P$ represents the elevation of the cell, while $E_S$ is defined as the average altitude of the surrounding pixels (Rahmati et al., 2018).

### ix. SPI

The stream position index (SPI) is used to assess the impact of water flow erosion on groundwater occurrence (Nampak et al., 2014). The following formula represents the equation for calculating SPI (Moore et al., 1991).

$$SPI = A_S \times tan\beta \quad (4)$$

The formula indicates $A_S$ as the flow accumulation of the watershed area and $\beta$ as the slope gradient (Fatah et al., 2024). Figure 3(i) illustrates the SPI map, which ranges from -15 to 14.

### x. Drainage Density

The relationship between groundwater and drainage density is inverse; low drainage density indicates high infiltration and less surface runoff, increasing groundwater level (Al-Kindi & Janizadeh, 2022; Dey et al., 2023). Figure 4(a) illustrates the drainage density map, which ranges from 0 to 4.

### xi. Lineament Density

Lineaments are linear or curved geological structures that signify faults and fractures formed through tectonic stress and geological processes (Dey et al., 2023; Jari et al., 2023). Lineament density was derived from the lineaments on the hillshade surface. Figure 4(b) illustrates the lineament density map, which ranges between 0 and 2.

### xii. Lithology

Lithology influences groundwater flow as it affects the aquifer permeability and soil porosity (Abrar et al., 2023; Islam et al., 2023; Jari et al., 2023; Maskooni et al., 2020). The Figure 4(c) illustrates ten classes of lithology map obtained from JMG.

### xiii. Geology

Geology affects the dynamics, retention, and capacity of groundwater (Naghibi et al., 2016). The permeability and porosity of the rock structure in geological formation will increase the infiltration rate towards groundwater occurrence (Dey et al., 2023). Figure 4(d) depicts difference classes of geological map of the Klang and Langat River basin.

### xiv. Distance to Fault

The data of major and minor faulting in the Klang and Langat River basin was acquired from JMG. Figure 4(e) illustrates the fault distance map produced using Euclidean Distance tools, with values ranging from 0 to 50472 m.

### xv. Distance to Streams

The distance to the stream map as illustrated in the Figure 4(f) was generated using Euclidean Distance tools with input derived from stream features obtained from LUAS. The map value ranges from 0 to 8035 m.

### xvi. Distance to Main Tributaries

The main tributaries feature was obtained from LUAS. The distance to main tributaries maps, illustrated in the Figure 4(g), was created with Euclidean Distance tools, with values ranging from 0 to 14168 m.

### xvii. LULC

LULC describes the existing geographic characteristics, including vegetation, soil cover, agricultural fields, urban infrastructure, and water bodies, which significantly affect groundwater movement (Ibrahim-Bathis & Ahmed, 2016; Shandu & Atif, 2023; Thanh et al., 2022). Figure 4(h) depicts the LULC map of the research region, which has five categories: water, forest, agriculture, developed land and barren land.

### xviii. Soil Types

Each soil type's characteristic gives a different permeability and retention rate towards groundwater occurrence (P. Kumar et al., 2016; Martínez-Santos et al., 2021; Rahmati et al., 2016). Soil also controls the infiltration rate, which determine how much water seeps through the subsurface layer (Seifu et al., 2023). The soil map in Figure 4(i) shows several classes of soil in the Klang and Langat River basin.

### xix. NDVI

NDVI as shown in Figure 5(a), influences groundwater by reflecting the characteristics of vegetation cover (Ragragui et al., 2024). Higher NDVI signifies robust vegetation, while lower NDVI denotes compromised vegetation. The following formula depicts the calculation of NDVI.

$$NDVI = \frac{NIR - R}{NIR + R} \quad (5)$$

NIR indicates near infrared, while R defines red, which represents the reflectance of vegetation (Prasad et al., 2020).

### xx. Rainfall

Rainfall influences groundwater recharge through precipitation (Prasad et al., 2020; Seifu et al., 2023; Thapa et al., 2017). Precipitation provides a primary source of water that percolates through soils and influences water recharge (Islam et al., 2023; Oikonomidis et al., 2015; Thapa et al., 2017). Figure 5(b) shows the rainfall map derived using IDW tools, which ranges between 73mm and 347mm.

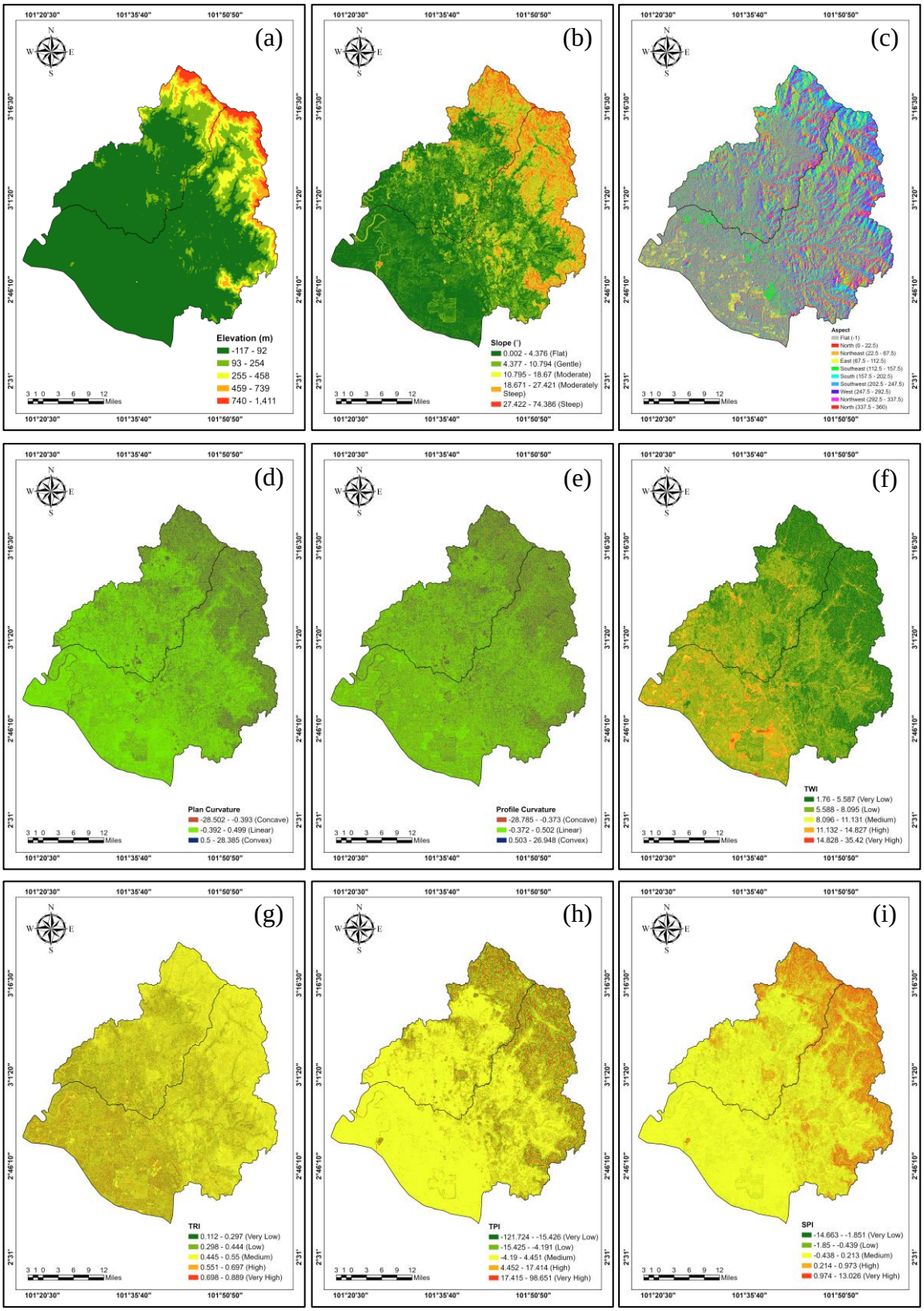


Figure 3: Thematic Layers for Groundwater Conditioning Parameters: (a) Elevation, (b) Slope, (c) Aspect, (d) Plan Curvature, (e) Profile Curvature, (f) TWI, (g) TRI, (h) TPI, (i) SPI

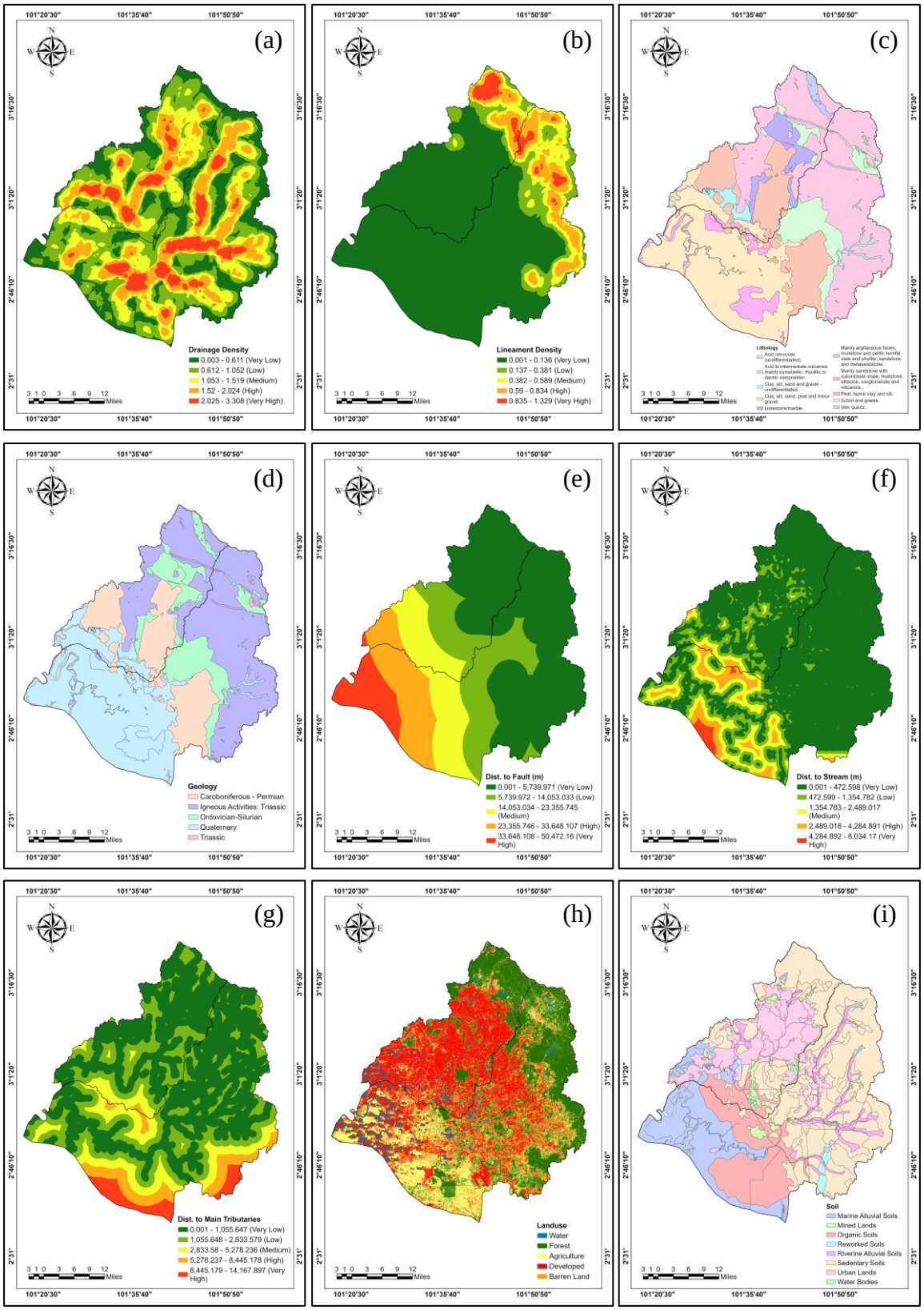


Figure 4: Thematic Layers for Groundwater Conditioning Parameters: (a) Drainage Density, (b) Lineament Density, (c) Lithology, (d) Geology, (e) Distance to Fault, (f) Distance to Streams, (g) Distance to Main Tributaries, (h) LULC, (i) Soil Types

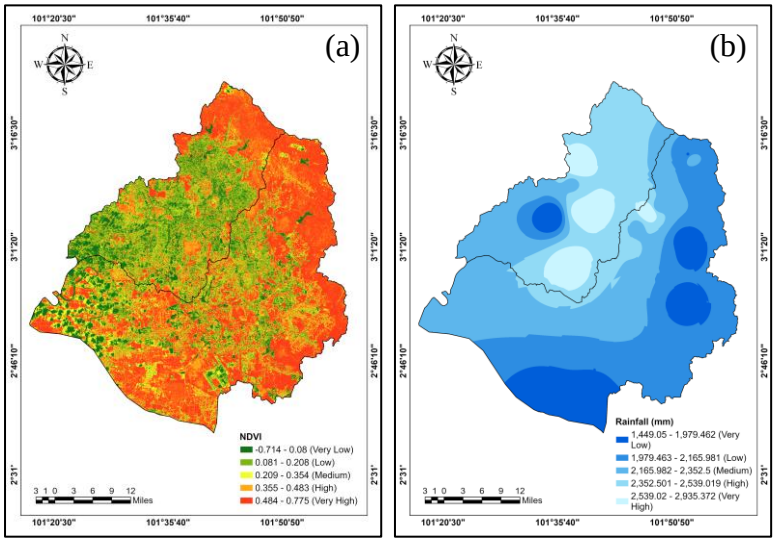


Figure 5: Thematic Layers for Groundwater Conditioning Parameters: (a) NDVI, (b) Rainfall

### c. GWP Parameters Optimization

#### i. Correlation Analysis

Correlation matrix is one of the methods that has been widely used by past studies to delineate groundwater influencing factors. Correlation matrix is an approach applied to ascertain the linear relationship among groundwater variables (Chouhan et al., 2024; Ejaz et al., 2024). The correlation analysis result ranges from -1 to 1, indicating the significance and direction of the linear connection between variables. A positive correlation signifies a strong association between variables, zero denotes a neutral relationship, and a negative correlation reflects an inverse relationship (K. Halder et al., 2024).

#### ii. Multicollinearity Test

The application of multicollinearity test will eliminate the possible noise error that might affect the prediction ability of the GWP model (S. Halder et al., 2024; Roy et al., 2024). Multicollinearity emerges when more than one variable in a model demonstrates a significant correlation (Ghosh & Bera, 2024; Roy et al., 2024). Two common indicators used by past studies to determine the collinearity issues which are the Variance Inflation Factor (VIF) and Tolerance (TOL) (Moghaddam et al., 2020; Ragragui et al., 2024; Razavi-Termeh et al., 2024; Wei et al., 2022). The following formula calculates the indicators (Anh et al., 2023; Ghosh & Bera, 2024).

$$TOL = 1 - R_V^2 \quad (6)$$

$$VIF = \frac{1}{TOL} \quad (7)$$

$R_V^2$ represents the coefficient of conditioning factors relative to all other groundwater conditioning factors (Anh et al., 2023). The VIF score must not above 10, and the TOL cannot exceed 0.1 to signify the absence of multicollinearity issues (Anh et al., 2023; Arabameri et al., 2019; Ghosh & Bera, 2024; Mukherjee & Singh, 2020). If the value is not within the threshold, the factors must be eliminated from the groundwater prediction model.

### iii. Chi-square Test

The chi-square test is used to measure significant relationships among variables (Pradhan et al., 2017). This study employed this method to evaluate the relationship between groundwater conditioning parameters and the target variables (tubewell points) (Liu et al., 2022). The subsequent formula represents the equation for the chi-square test.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad (8)$$

The formula defines $\chi^2$ as chi-squared, $O_i$ as observed value and $E_i$ as expected value. A higher chi-square ($\chi^2$) value contributes to the significant relationship between variables, whereas a lower value suggests no correlation between the target and variables (Ganesh et al., 2025; Liu et al., 2022). The p-value in this test were assessed to determine the statistically significant contribution towards GWP prediction. A p-value below than 0.05 indicates as significant contribution towards GWP, while p-value above 0.05 suggests no significant influence (Ganesh et al., 2025).

## d. Random Forest (RF)

RF is a widely used algorithms in machine learning, introduced by Breiman in 2001. Random forest is an ensemble method that merges many decision trees to enhance accuracy and decrease the probability of overfitting. Random Forest may be used for tasks involving regression and classification. Previous studies (Djerida et al., 2024; K. Halder et al., 2024; Z. A. Khan & Jhamnani, 2023; R. Kumar et al., 2021; Seifu et al., 2023) have demonstrated that RF is constantly exhibits higher performance for AUC(ROC) and overall accuracy.

## e. Evaluation Metrics

Evaluation metrics are essential in machine learning to measure the model capability and performance. This study used several assessment measures, including accuracy, precision,

recall, F1-score, AUC, and ROC curve, to assess the prediction performance of the RF model. As shown in Equation 9, accuracy indicates as the ratio of correct predictions produced by the model compared to the total predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

Precision (Equation 10) quantifies the ratio of accurately predicted groundwater locations to all positive predictions, while recall (Equation 11) assesses the model's capacity to identify the actual groundwater occurrences.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

The F1-score in Equation 12 represents the harmonic mean of accuracy and recall, offering an equal evaluation of the two measures.

$$F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (12)$$

The equations above denote:

- $TP$: True positive
- $TN$: True negative
- $FP$: False positive
- $FN$: False negative.

Finally, the area under curve (AUC) and receiver operating characteristics (ROC) curve were used to evaluate the model's discriminative capabilities between tubewell and non-tubewell locations. In ROC curve, the true positive rate (TPR) at different categorization criteria is shown in relation to the false positive rate (FPR). AUC, on the other hand, evaluates the model's overall capacity to differentiate between positive and negative classifications. An AUC value of 0 indicates that the model performs worse than random guessing, while an AUC value of 1 indicates that the model successfully separates positive and negative classifications.

**Results and Discussion**

**a. Analysis of GWP Parameters**

## i. Correlation Matrix

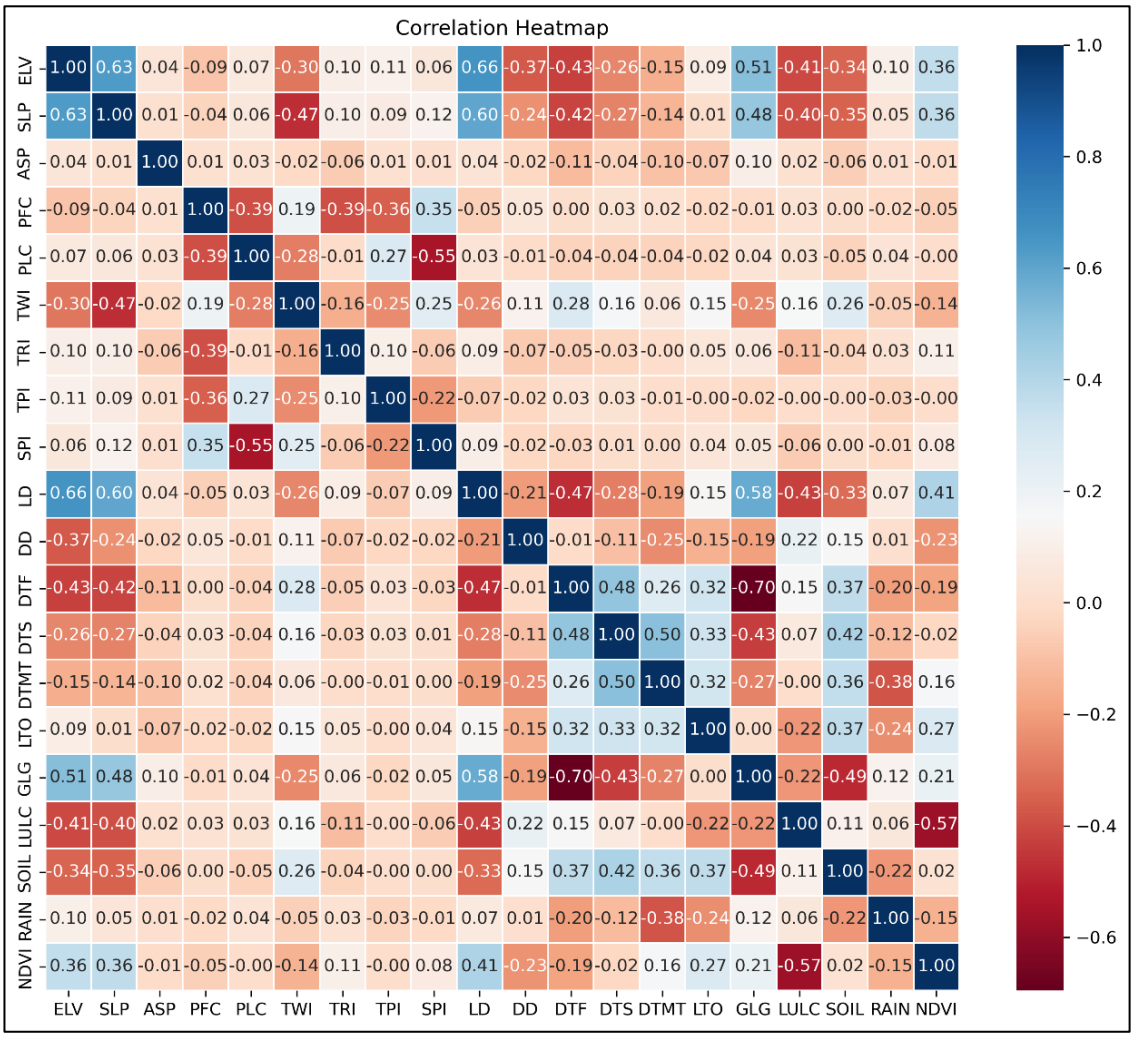


Figure 6: Correlation Matrix

Figure 6 illustrates the results derived from the correlation study of groundwater conditioning parameters. The correlation result signifies a value around -1 indicates a strong negative association, whereas a number near +1 indicates a strong positive association. Based on Figure 6, most of the groundwater conditioning parameters exhibit a weak to moderate relationship, which suggests minimal redundancy among variables and reduces the risk of multicollinearity in the modelling process. The strongest positive relationships were found between elevation (ELV) and lineament density (LD) (r = 0.67), followed by elevation (ELV) and slope (SLP) (r = 0.59). This indicated that higher elevation was strongly associated with steeper slopes and higher lineaments, which are important for groundwater occurrence. In contrast, the strongest negative relationship was found between geology (GLG) and distance to fault (DTF), with a value of –0.71, indicating that specific geology attributes were strongly correlated with fault structures that determine groundwater

occurrence. Thus, the results from the correlation analysis indicate that all variables were accepted and do not exceed the commonly used threshold of 0.8.

### ii. Multicollinearity Test

Table 1: Multicollinearity Test Result

| Features | VIF | TOL |
|---|---|---|
| Elevation | 2.516 | 0.397 |
| Slope | 2.372 | 0.422 |
| Aspect | 1.036 | 0.965 |
| Profile Curvature | 1.612 | 0.620 |
| Plan Curvature | 1.659 | 0.603 |
| TWI | 1.613 | 0.620 |
| TRI | 1.275 | 0.784 |
| TPI | 1.283 | 0.779 |
| SPI | 1.605 | 0.623 |
| Drainage Density | 1.461 | 0.685 |
| Lineament Density | 2.539 | 0.394 |
| Lithology | 2.051 | 0.488 |
| Geology | 3.213 | 0.311 |
| Distance to Fault | 3.112 | 0.321 |
| Distance to Stream | 1.805 | 0.554 |
| Distance to Main Tributaries | 1.893 | 0.528 |
| Soil | 1.886 | 0.530 |
| LULC | 1.670 | 0.588 |
| NDVI | 1.794 | 0.557 |
| Rainfall | 1.294 | 0.773 |

Table 1 shows the result of the multicollinearity test for groundwater conditioning parameters, the VIF values range from 1.050 to 3.157, while TOL values range from 0.317 to 0.952. These results indicate that there are no variables that exceed the required threshold (VIF < 10; TOL > 0.10), which suggests that no multicollinearity issues arise among the variables. Therefore, all of the groundwater conditioning parameters were accepted and can be used in the modelling processes, which is consistent with the previous studies by Anh et al. (2023), Ghosh & Bera (2024) and Moghaddam et al. (2020).

### iii. Chi Square Test

Table 2: Chi-Square and p-value Result

| Features | Chi – Square ($\chi^2$) | $\rho$ – value |
|---|---|---|
| Lineament Density | 67.259 | 0.000 |
| Elevation | 51.526 | 0.000 |
| Geology | 46.222 | 0.000 |
| Soil | 33.195 | 0.000 |
| Slope | 29.412 | 0.000 |
| Distance to Fault | 22.721 | 0.000 |
| LULC | 18.230 | 0.000 |
| NDVI | 17.166 | 0.000 |
| Drainage Density | 11.997 | 0.000 |
| Distance to Stream | 3.660 | 0.056 |
| TWI | 3.131 | 0.077 |
| Distance to Main Tributaries | 1.858 | 0.173 |
| TRI | 1.101 | 0.294 |
| Aspect | 0.860 | 0.354 |
| Lithology | 0.695 | 0.405 |
| Rainfall | 0.471 | 0.493 |
| Profile Curvature | 0.079 | 0.778 |
| TPI | 0.038 | 0.846 |
| SPI | 0.018 | 0.893 |
| Plan Curvature | 0.010 | 0.919 |

Table 2 shows the chi-square ($\chi^2$) and p-value results between tubewell distribution and groundwater conditioning parameters. Based on the result, nine discovered variables (lineament density, elevation, geology, soil, slope, distance to fault, land use/land cover, NDVI, and drainage density) have a p-value < 0.05, indicating statistical significance for groundwater occurrence. It is supported by the higher $\chi^2$ value that indicates these parameters highly influenced GWP prediction. In contrast, the remaining parameters, such as distance to stream, TWI, distance to main tributaries, TRI, aspect, lithology, rainfall, profile curvature, TPI, SPI and plan curvature, demonstrate a p-value > 0.05, signifying no statistically significant impact on groundwater occurrence. Therefore, nine parameters highly influenced towards groundwater occurrence were retained for further modelling, while the less influenced parameters were excluded to enhance the reliability of the GWP prediction.

**b. GWP Using Random Forest**

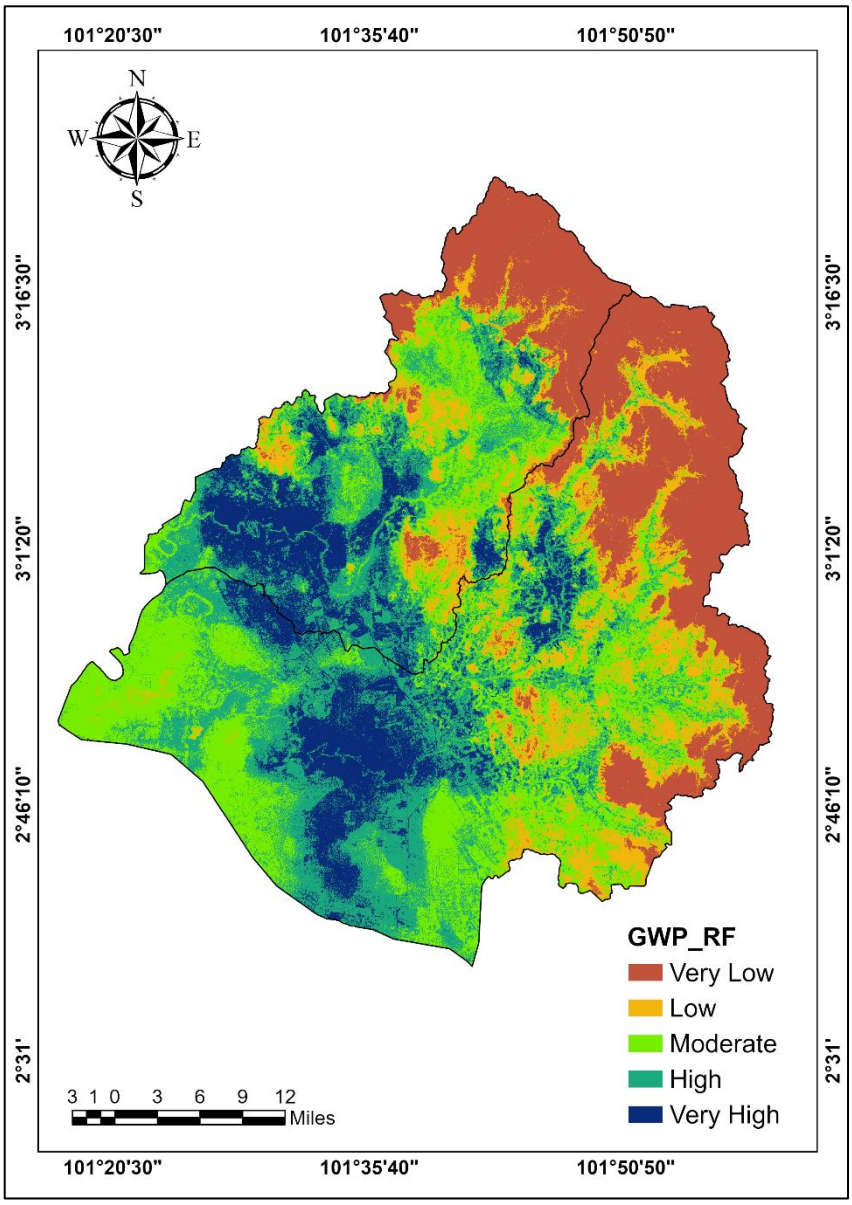


Figure 7: GWP Area Using RF

Figure 7 depicts the groundwater potential map derived from the RF model. The GWP map was sorted into five classifications: very low, low, moderate, high and very high. The areas with the largest anticipated groundwater potential were mostly situated in the centre portion of the research area, shown in dark blue. These areas were likely influenced by gentle slope, agricultural land, high rainfall distribution, and lithology dominated by clay, silt, sand, peat, and minor gravel. These factors enhanced the infiltration and storage capacity, which led to groundwater occurrence (Anh et al., 2023; Fatah et al., 2024; Ghosh & Bera, 2024; Islam et al., 2023). Meanwhile, the areas with the lowest groundwater potential were situated in the northern and northeastern sections of the research area, shown in red. These areas were influenced by several factors such as steep slopes, higher elevations, dense forest cover and

lithology dominated by vein quartz types, limiting infiltration and storage capacity (Al-Kindi & Janizadeh, 2022; Manap et al., 2014; Masroor et al., 2021).

Table 3: GWP Area Percentages

| Classes | Area ($km^2$) | Percentages |
|---|---|---|
| Very High | 541.17 | 14.02 |
| High | 976.29 | 25.30 |
| Moderate | 1055.29 | 27.35 |
| Low | 492.00 | 12.75 |
| Very Low | 794.38 | 20.58 |

Table 3 presents the area and percentages corresponding to each groundwater potential classification. The results demonstrate the moderate class area has the highest predicted area (1055.29 $km^2$), followed by the high class (976.29 $km^2$). Meanwhile, the low class exhibited the lowest predicted area (492 $km^2$). Consequently, the findings suggest that the majority of the research area is situated within moderate to high groundwater potential zones, with only limited zones exhibiting low groundwater potential.

### c. Evaluation Metrics of GWP Using RF

Table 4: Evaluation Metrics Result for Training and Testing

| Evaluation Metrics | Training | Testing |
|---|---|---|
| Accuracy | 0.824 | 0.743 |
| Precision | 0.786 | 0.724 |
| Recall | 0.892 | 0.786 |
| F1 Score | 0.836 | 0.754 |
| ROC(AUC) | 0.927 | 0.860 |

Table 4 depicts the evaluation metrics for the training and testing datasets obtained from GWP prediction. The model demonstrated strong predictive performance by achieving 0.824 for accuracy on training and 0.743 for testing, indicating effective generalization with only a moderate decreased performance on unseen data. Precision and recall declined from 0.786 to 0.724 and from 0.892 to 0.786 for the training and testing datasets, respectively. The results suggest that the model is more effective in correctly identifying groundwater locations than reducing false positives. Meanwhile, the F1-score remained consistently high, at 0.836 for training and 0.754 for testing, indicating a balanced performance between precision and recall.

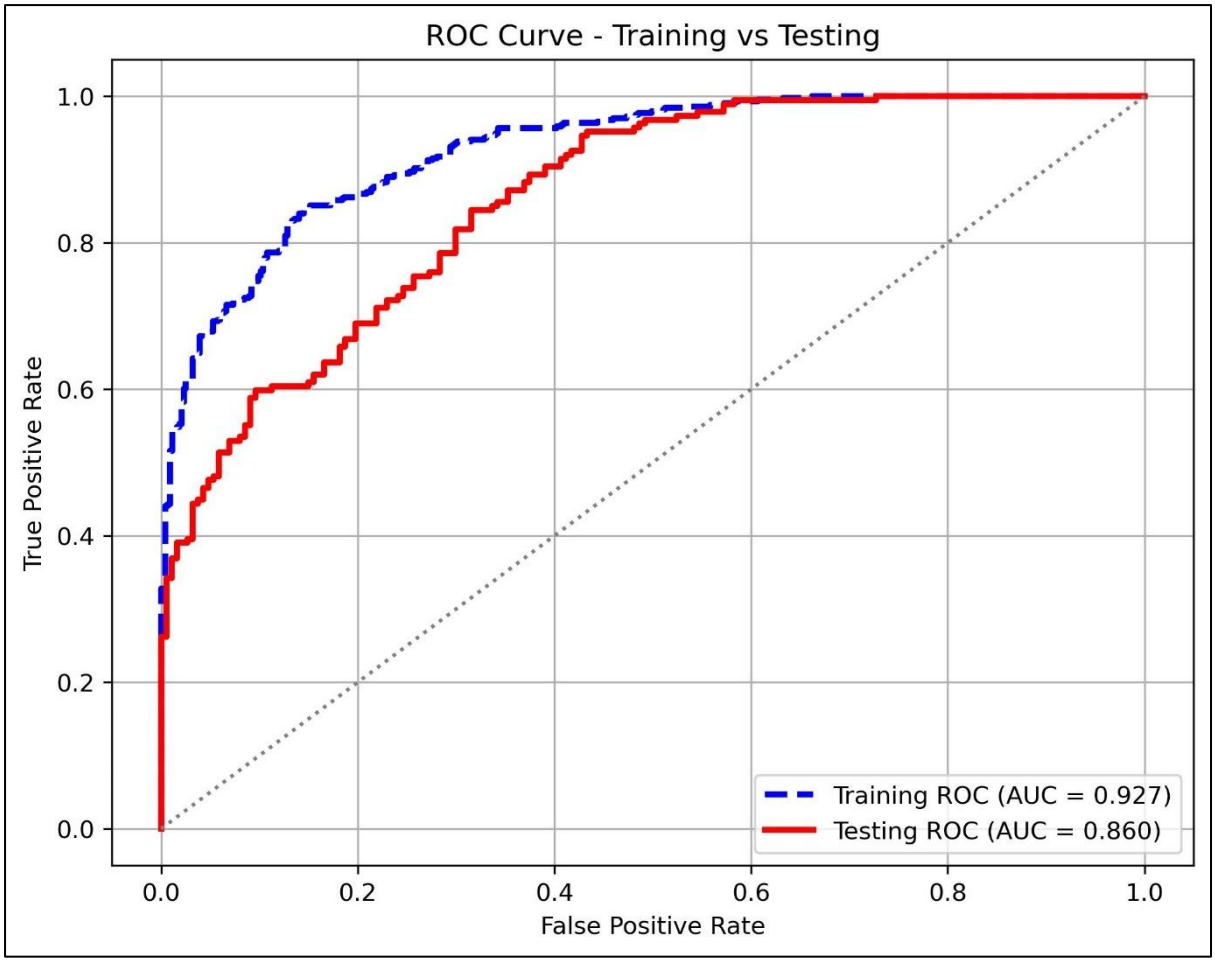


Figure 8: AUC and ROC Curve

Figure 8 depicts the AUC and ROC curve obtained from the RF models. The AUC value of 0.927 for the training dataset implies an excellent classification performance, while the slightly lower value of 0.860 for the testing dataset indicates a strong generalization to the unseen data. The difference between the training and testing data for AUC demonstrates that the model was not significantly overfitted and maintains the capability of reliable predictive modelling for groundwater potential. The result of this study is aligned with the studies from Arabameri et al. (2019), Das & Saha (2022) and Maskooni et al. (2020), which demonstrated an AUC value greater than 0.80 that reflecting to a good and excellent result of GWP model.

**Conclusion and Recommendation**

This research optimized and enhanced GWP parameters using several statistical approaches: correlation analysis, multicollinearity test and chi-square test. Based on the correlation analysis result, most of the GWP parameters exhibit minimal redundancy, while elevation shows a stronger relationship between elevation and lineament density (r=0.67) and between elevation and slope (r=0.59). For the multicollinearity test result, all of the parameters were

within the required threshold for both VIF (<10) and TOL (>0.10). This result indicates there are no collinearity issues among variables. However, according to the chi-square results, nine variables: lineament density, elevation, geology, soil, slope, distance to fault, LULC, NDVI and drainage density, significantly contributed to groundwater occurrence, while the remaining of the parameters give minimal contributions and were removed to enhance the reliability of the model. These findings emphasize the efficacy of optimization analysis in distinguishing the most critical variables that would contribute to the groundwater identification and removing variables with zero predicting abilities that might affect the model's accuracy and reliability. The optimized GWP model using RF achieved excellent performance of AUC with the value of 0.927 for training and 0.860 for testing. In conclusion, integrating optimized parameters with machine learning models provides a highly reliable and robust framework for GWP modelling. Further studies should be recommended to apply a similar or more advanced optimization method that could remove the redundancy, prediction errors and enhance the model. Also, integrating other factors such as subsurface data or ground truth could improve the groundwater potential assessment's accuracy, precision and reliability for sustainable groundwater resource management.

**Acknowledgements**

We would like to thank Department of Mineral and Geoscience (JMG) Malaysia, Department of Agriculture Malaysia, Malaysia Meteorological Department (METMalaysia) and Department of Drainage and Irrigation for providing the hydrogeological, soil and rainfall data. We are also grateful to The Ministry of Higher Education for the funding provided under grant number (Grant no. FRGS/1/2023/WAB07/UITM/02/3) and the Universiti Teknologi MARA.

**References**

Abrar, H., Legesse Kura, A., Esayas Dube, E., & Likisa Beyene, D. (2023). AHP based analysis of groundwater potential in the western escarpment of the Ethiopian rift valley. *Geology, Ecology, and Landscapes*, *7*(3), 175–188. https://doi.org/10.1080/24749508.2021.1952761

Al-Kindi, K. M., & Janizadeh, S. (2022). Machine Learning and Hyperparameters Algorithms for Identifying Groundwater Aflaj Potential Mapping in Semi-Arid Ecosystems Using LiDAR, Sentinel-2, GIS Data, and Analysis. *Remote Sensing*, *14*(21). https://doi.org/10.3390/rs14215425

Anh, D. T., Pandey, M., Mishra, V. N., Singh, K. K., Ahmadi, K., Janizadeh, S., Tran, T. T., Linh, N. T. T., & Dang, N. M. (2023). Assessment of groundwater potential modeling using support vector machine optimization based on Bayesian multi-objective hyperparameter algorithm. *Applied Soft Computing*, *132*. https://doi.org/10.1016/j.asoc.2022.109848

Arabameri, A., Rezaei, K., Cerda, A., Lombardo, L., & Rodrigo-Comino, J. (2019). GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Science of the Total Environment*, *658*, 160–177. https://doi.org/10.1016/j.scitotenv.2018.12.115

Azizan, F. A., Aznan, A. A., Ruslan, R., Nazari, M., & Jaafar, M. N. (2018). Groundwater Assessment Using Geophysical Survey at Insat, Perlis, Malaysia. *IOP Conference Series: Materials Science and Engineering*, *429*(1). https://doi.org/10.1088/1757-899X/429/1/012026

Breiman, L. (2001). *Random Forests* (Vol. 45).

Chouhan, A. K., Harsh, A., Mishra, A. K., Kumar, V., Kumar, R., & Kumar, S. (2024). Delineation of groundwater vulnerable zone for sustainable development in the southwestern part of Bihar, India. *Groundwater for Sustainable Development*, *26*. https://doi.org/10.1016/j.gsd.2024.101240

Dahal, K., Sharma, S., Shakya, A., Talchabhadel, R., Adhikari, S., Pokharel, A., Sheng, Z., Pradhan, A. M. S., & Kumar, S. (2023). Identification of groundwater potential zones in data-scarce mountainous region using explainable machine learning. *Journal of Hydrology*, *627*. https://doi.org/10.1016/j.jhydrol.2023.130417

Das, R., & Saha, S. (2022). Spatial mapping of groundwater potentiality applying ensemble of computational intelligence and machine learning approaches. *Groundwater for Sustainable Development*, *18*. https://doi.org/10.1016/j.gsd.2022.100778

Dey, B., Abir, K. A. M., Ahmed, R., Salam, M. A., Redowan, M., Miah, M. D., & Iqbal, M. A. (2023). Monitoring groundwater potential dynamics of north-eastern Bengal Basin in Bangladesh using AHP-Machine learning approaches. *Ecological Indicators*, *154*. https://doi.org/10.1016/j.ecolind.2023.110886

Djerida, A., Bennia, A., & Kebir, L. W. (2024). Application of deep neural networks to groundwater potential mapping in the region of Sidi Bel Abbès, Algeria. *2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium, M2GARSS 2024 - Proceedings*, 463–467. https://doi.org/10.1109/M2GARSS57310.2024.10537597

Ejaz, N., Khan, A. H., Saleem, M. W., Elfeki, A. M., Rahman, K. U., Hussain, S., Ullah, S., & Shang, S. (2024). Multi-criteria decision-making techniques for groundwater potentiality mapping in arid regions: A case study of Wadi Yiba, Kingdom of Saudi Arabia. *Groundwater for Sustainable Development*, *26*. https://doi.org/10.1016/j.gsd.2024.101223

El Shinawi, A., Kuriqi, A., Zelenakova, M., Vranayova, Z., & Abd-Elaty, I. (2022). Land subsidence and environmental threats in coastal aquifers under sea level rise and over-

pumping stress. *Journal of Hydrology*, *608*. https://doi.org/10.1016/j.jhydrol.2022.127607

Fatah, K. K., Mustafa, Y. T., & Hassan, I. O. (2024). Groundwater potential mapping in arid and semi-arid regions of kurdistan region of Iraq: A geoinformatics-based machine learning approach. *Groundwater for Sustainable Development*, *27*. https://doi.org/10.1016/j.gsd.2024.101337

Ganesh, B., Vincent, S., Pathan, S., Bhat, G. V., & Bhat K, J. (2025). Optimized Machine Learning based Model for the Large-Scale Spatial Prediction of Landslides at Western Ghats in the state of Karnataka, India. *Natural Hazards Research*. https://doi.org/10.1016/j.nhres.2025.08.004

Ghosh, A., & Bera, B. (2024). Potentialities and development of groundwater resources applying machine learning models in the extended section of Manbhum-Singhbhum Plateau, India. *HydroResearch*, *7*, 1–14. https://doi.org/10.1016/j.hydres.2023.11.002

Halder, K., Srivastava, A. K., Ghosh, A., Nabik, R., Pan, S., Chatterjee, U., Bisai, D., Pal, S. C., Zeng, W., Ewert, F., Gaiser, T., Pande, C. B., Islam, A. R. M. T., Alam, E., & Islam, M. K. (2024). Application of bagging and boosting ensemble machine learning techniques for groundwater potential mapping in a drought-prone agriculture region of eastern India. *Environmental Sciences Europe*, *36*(1). https://doi.org/10.1186/s12302-024-00981-y

Halder, S., Karmakar, S., Maiti, P., Roy, M. B., & Roy, P. K. (2024). Application of machine learning and fuzzy AHP for identification of suitable groundwater potential zones using field based hydrogeophysical and soil hydraulic factors in a complex hydrogeological terrain. *Groundwater for Sustainable Development*, *27*. https://doi.org/10.1016/j.gsd.2024.101329

Ibrahim-Bathis, K., & Ahmed, S. A. (2016). Geospatial technology for delineating groundwater potential zones in Doddahalla watershed of Chitradurga district, India. *Egyptian Journal of Remote Sensing and Space Science*, *19*(2), 223–234. https://doi.org/10.1016/j.ejrs.2016.06.002

Islam, F., Tariq, A., Guluzade, R., Zhao, N., Shah, S. U., Ullah, M., Hussain, M. L., Ahmad, M. N., Alasmari, A., Alzuaibr, F. M., Askary, A. El, & Aslam, M. (2023). Comparative analysis of GIS and RS based models for delineation of groundwater potential zone mapping. *Geomatics, Natural Hazards and Risk*, *14*(1). https://doi.org/10.1080/19475705.2023.2216852

Jari, A., Bachaoui, E. M., Hajaj, S., Khaddari, A., Khandouch, Y., El Harti, A., Jellouli, A., & Namous, M. (2023). Investigating machine learning and ensemble learning models in groundwater potential mapping in arid region: case study from Tan-Tan water-scarce region, Morocco. *Frontiers in Water*, *5*. https://doi.org/10.3389/frwa.2023.1305998

Kalantar, B., Al-Najjar, H. A. H., Pradhan, B., Saeidi, V., Halin, A. A., Ueda, N., & Naghibi, S. A. (2019). Optimized conditioning factors using machine learning techniques for groundwater potential mapping. *Water (Switzerland)*, *11*(9). https://doi.org/10.3390/w11091909

Khan, M. M. A., Raj, K., Rak, A. A. E., Mansor, H. E., Mostapa, R., Samuding, K., & Shah, Z. A. (2021). Stable isotope evidence on mechanisms and sources of groundwater recharge in quaternary aquifers of Kelantan, Malaysia. *Arabian Journal of Geosciences*, *14*(16). https://doi.org/10.1007/s12517-021-07646-7

Khan, Z. A., & Jhamnani, B. (2023). Identification of groundwater potential zones of Idukki district using remote sensing and GIS-based machine-learning approach. *Water Supply*, *23*(6), 2426–2446. https://doi.org/10.2166/ws.2023.134

Kumar, M., Singh, P., & Singh, P. (2023). Machine learning and GIS-RS-based algorithms for mapping the groundwater potentiality in the Bundelkhand region, India. *Ecological Informatics*, *74*. https://doi.org/10.1016/j.ecoinf.2023.101980

Kumar, P., Herath, S., Avtar, R., & Takeuchi, K. (2016). Mapping of groundwater potential zones in Killinochi area, Sri Lanka, using GIS and remote sensing techniques. *Sustainable Water Resources Management*, *2*(4), 419–430. https://doi.org/10.1007/s40899-016-0072-5

Kumar, P., Singh, P., Asthana, H., Yadav, B., & Mukherjee, S. (2024). Groundwater potential zone mapping of middle Andaman using multi-criteria decision-making and support vector machine. *Groundwater for Sustainable Development*, *26*. https://doi.org/10.1016/j.gsd.2024.101191

Kumar, R., Dwivedi, S. B., & Gaur, S. (2021). A comparative study of machine learning and Fuzzy-AHP technique to groundwater potential mapping in the data-scarce region. *Computers and Geosciences*, *155*. https://doi.org/10.1016/j.cageo.2021.104855

Liu, R., Li, G., Wei, L., Xu, Y., Gou, X., Luo, S., & Yang, X. (2022). Spatial prediction of groundwater potentiality using machine learning methods with Grey Wolf and Sparrow Search Algorithms. *Journal of Hydrology*, *610*. https://doi.org/10.1016/j.jhydrol.2022.127977

Manap, M. A., Nampak, H., Pradhan, B., Lee, S., Sulaiman, W. N. A., & Ramli, M. F. (2014). Application of probabilistic-based frequency ratio model in groundwater potential mapping using remote sensing data and GIS. *Arabian Journal of Geosciences*, *7*(2), 711–724. https://doi.org/10.1007/s12517-012-0795-z

Martínez-Santos, P., Aristizábal, H. F., Díaz-Alcaide, S., & Gómez-Escalonilla, V. (2021). Predictive mapping of aquatic ecosystems by means of support vector machines and random forests. *Journal of Hydrology*, *595*(June 2020). https://doi.org/10.1016/j.jhydrol.2021.126026

Maskooni, E. K., Naghibi, S. A., Hashemi, H., & Berndtsson, R. (2020). Application of advanced machine learning algorithms to assess groundwater potential using remote sensing-derived data. *Remote Sensing*, *12*(17). https://doi.org/10.3390/RS12172742

Masroor, M., Rehman, S., Sajjad, H., Rahaman, M. H., Sahana, M., Ahmed, R., & Singh, R. (2021). Assessing the impact of drought conditions on groundwater potential in Godavari Middle Sub-Basin, India using analytical hierarchy process and random forest machine learning algorithm. *Groundwater for Sustainable Development*, *13*. https://doi.org/10.1016/j.gsd.2021.100554

Moghaddam, D. D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A. T., Nalivan, O. A., & Tien Bui, D. (2020). The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *Catena*, *187*. https://doi.org/10.1016/j.catena.2019.104421

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). DIGITAL TERRAIN MODELLING: A REVIEW zy OF HYDROLOGICAL, GEOMORPHOLOGICAL, AND BIOLOGICAL APPLICATIONS. In *HYDROLOGICAL PROCESSES* (Vol. 5).

Mukherjee, I., & Singh, U. K. (2020). Delineation of groundwater potential zones in a drought-prone semi-arid region of east India using GIS and analytical hierarchical process techniques. *Catena*, *194*. https://doi.org/10.1016/j.catena.2020.104681

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, *188*(1), 1–27. https://doi.org/10.1007/s10661-015-5049-6

Nampak, H., Pradhan, B., & Manap, M. A. (2014). Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. *Journal of Hydrology*, *513*, 283–300. https://doi.org/10.1016/j.jhydrol.2014.02.053

Oikonomidis, D., Dimogianni, S., Kazakis, N., & Voudouris, K. (2015). A GIS/Remote Sensing-based methodology for groundwater potentiality assessment in Tirnavos area, Greece. *Journal of Hydrology*, *525*, 197–208. https://doi.org/10.1016/j.jhydrol.2015.03.056

Ouali, L., Kabiri, L., Namous, M., Hssaisoune, M., Abdelrahman, K., Fnais, M. S., Kabiri, H., El Hafyani, M., Oubaassine, H., Arioua, A., & Bouchaou, L. (2023). Spatial Prediction of Groundwater Withdrawal Potential Using Shallow, Hybrid, and Deep Learning Algorithms in the Toudgha Oasis, Southeast Morocco. *Sustainability (Switzerland)*, *15*(5). https://doi.org/10.3390/su15053874

Panneerselvam, B., Muniraj, K., Duraisamy, K., Pande, C., Karuppannan, S., & Thomas, M. (2023). An integrated approach to explore the suitability of nitrate-contaminated groundwater for drinking purposes in a semiarid region of India. *Environmental Geochemistry and Health*, *45*(3), 647–663. https://doi.org/10.1007/s10653-022-01237-5

Pradhan, B., Seeni, M. I., & Kalantar, B. (2017). Performance evaluation and sensitivity analysis of expert-based, statistical, machine learning, and hybrid models for producing landslide susceptibility maps. In *Laser Scanning Applications in Landslide Assessment* (pp. 193–232). Springer International Publishing. https://doi.org/10.1007/978-3-319-55342-9_11

Prasad, P., Loveson, V. J., Kotha, M., & Yadav, R. (2020). Application of machine learning techniques in groundwater potential mapping along the west coast of India. *GIScience and Remote Sensing*, 735–752. https://doi.org/10.1080/15481603.2020.1794104

Ragragui, H., Aouragh, M. H., El-Hmaidi, A., Ouali, L., Saouita, J., Iallamen, Z., Ousmana, H., Jaddi, H., & El Ouali, A. (2024). Mapping and modeling groundwater potential using machine learning, deep learning and ensemble learning models in the Saiss basin (Fez-

Meknes region, Morocco). *Groundwater for Sustainable Development*, *26*. https://doi.org/10.1016/j.gsd.2024.101281

Rahmati, O., Naghibi, S. A., Shahabi, H., Bui, D. T., Pradhan, B., Azareh, A., Rafiei-Sardooi, E., Samani, A. N., & Melesse, A. M. (2018). Groundwater spring potential modelling: Comprising the capability and robustness of three different modeling approaches. *Journal of Hydrology*, *565*, 248–261. https://doi.org/10.1016/j.jhydrol.2018.08.027

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *Catena*, *137*, 360–372. https://doi.org/10.1016/j.catena.2015.10.010

Razavi-Termeh, S. V., Sadeghi-Niaraki, A., Abba, S. I., Ali, F., & Choi, S. M. (2024). Enhancing spatial prediction of groundwater-prone areas through optimization of a boosting algorithm with bio-inspired metaheuristic algorithms. *Applied Water Science*, *14*(11). https://doi.org/10.1007/s13201-024-02301-4

Roy, S. K., Hasan, M. M., Mondal, I., Akhter, J., Roy, S. K., Talukder, S., Islam, A. K. M. S., Rahman, A., & Karuppannan, S. (2024). Empowered machine learning algorithm to identify sustainable groundwater potential zone map in Jashore District, Bangladesh. *Groundwater for Sustainable Development*, *25*. https://doi.org/10.1016/j.gsd.2024.101168

Seifu, T. K., Eshetu, K. D., Woldesenbet, T. A., Alemayehu, T., & Ayenew, T. (2023). Application of advanced machine learning algorithms and geospatial techniques for groundwater potential zone mapping in Gambela Plain, Ethiopia. *Hydrology Research*, *54*(10), 1246–1266. https://doi.org/10.2166/nh.2023.083

Shandu, I. D., & Atif, I. (2023). An Integration of Geospatial Modelling and Machine Learning Techniques for Mapping Groundwater Potential Zones in Nelson Mandela Bay, South Africa. *Water (Switzerland)*, *15*(19). https://doi.org/10.3390/w15193447

Sharifi, A., Khodaei, B., Ahrari, A., Hashemi, H., & Torabi Haghighi, A. (2024). Can river flow prevent land subsidence in urban areas? *Science of the Total Environment*, *917*. https://doi.org/10.1016/j.scitotenv.2024.170557

Sharma, Y., Ahmed, R., Saha, T. K., Bhuyan, N., Kumari, G., Roshani, Pal, S., & Sajjad, H. (2024). Assessment of groundwater potential and determination of influencing factors using remote sensing and machine learning algorithms: A study of Nainital district of Uttarakhand state, India. *Groundwater for Sustainable Development*, *25*. https://doi.org/10.1016/j.gsd.2024.101094

Sørensen, R., Zinko, U., & Seibert, J. (2006). On the calculation of the topographic wetness index: Evaluation of different methods based on field observations. *Hydrology and Earth System Sciences*, *10*(1), 101–112. https://doi.org/10.5194/hess-10-101-2006

Thanh, N. N., Chotpantarat, S., Trung, N. H., Ngu, N. H., & Muoi, L. Van. (2022). Mapping groundwater potential zones in Kanchanaburi Province, Thailand by integrating of analytic hierarchy process, frequency ratio, and random forest. *Ecological Indicators*, *145*. https://doi.org/10.1016/j.ecolind.2022.109591

Thapa, R., Gupta, S., Guin, S., & Kaur, H. (2017). Assessment of groundwater potential zones using multi-influencing factor (MIF) and GIS: a case study from Birbhum district, West Bengal. *Applied Water Science*, *7*(7), 4117–4131. https://doi.org/10.1007/s13201-017-0571-z

Wei, A., Li, D., Bai, X., Wang, R., Fu, X., & Yu, J. (2022). Application of machine learning to groundwater spring potential mapping using averaging, bagging, and boosting techniques. *Water Supply*, *22*(8), 6882–6894. https://doi.org/10.2166/ws.2022.283